

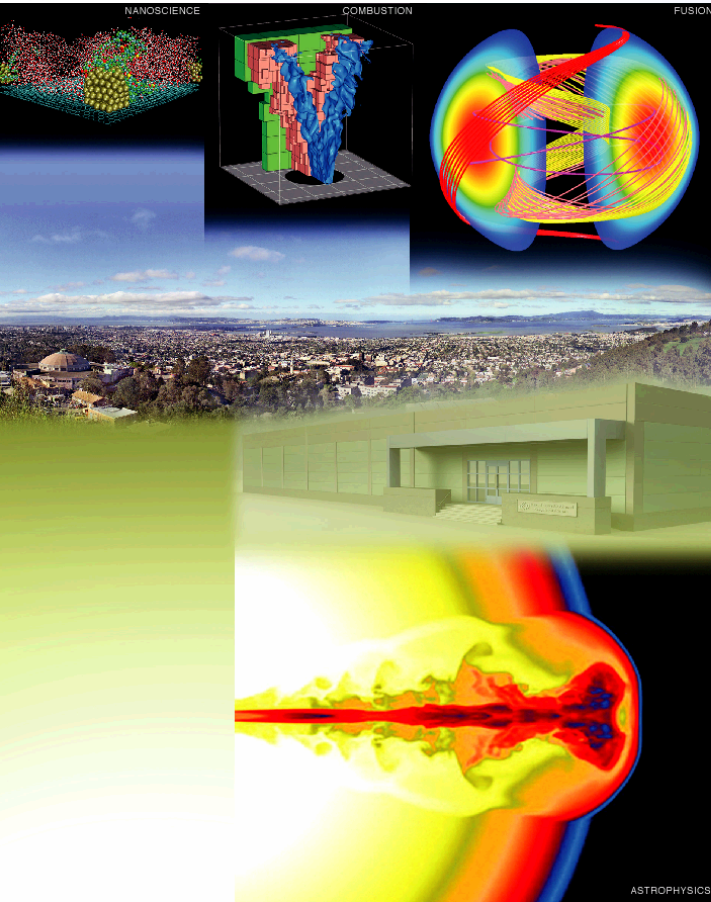


NERSC and HPCMP Cooperation

William T. Kramer
NERSC/LBNL
June 29, 2005



Outline



- Introduction to NERSC
- Motivation for collaboration with HPCMP
- Experiences in coordinated procurement
- Potential areas of further collaboration



NERSC Mission

The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate the pace of scientific discovery by providing high performance computing, information, data, and communications services for research sponsored by the DOE Office of Science (SC).

- **Support open, unclassified, basic research**
- **Deliver a complete environment (computing, storage, visualization, networking, grid services, cybersecurity)**
- **Focus on intellectual services to enable computational science**
- **Close collaborations between UC and NERSC in computer science and computational science**

National Energy Research Scientific Computing Center

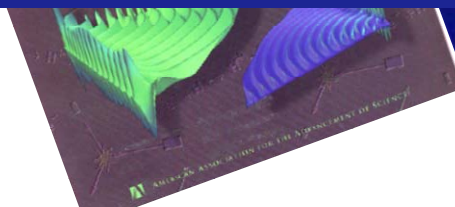
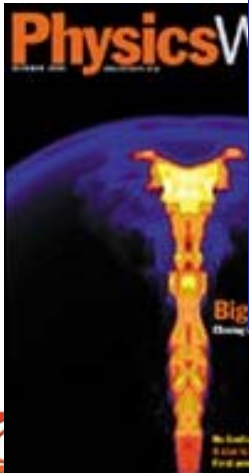
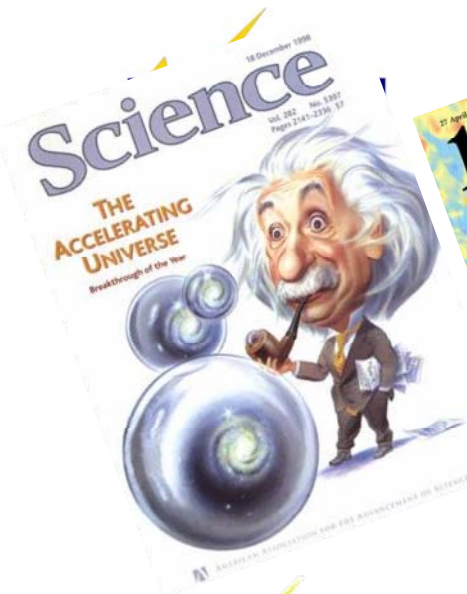
Serves the entire scientific
community

~2500 Users in
~250 projects

- Focus on
large-scale
computing

- In 2003, NERSC users alone reported the publication of at least 2,404 papers that were partly based on work done at NERSC.

- In 2004, NERSC users reported the publication of at least 1,100 papers that were partly based on work done at NERSC

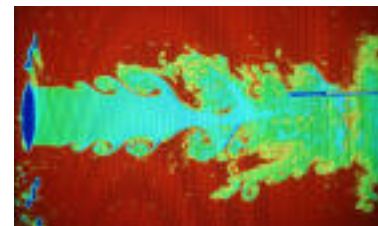
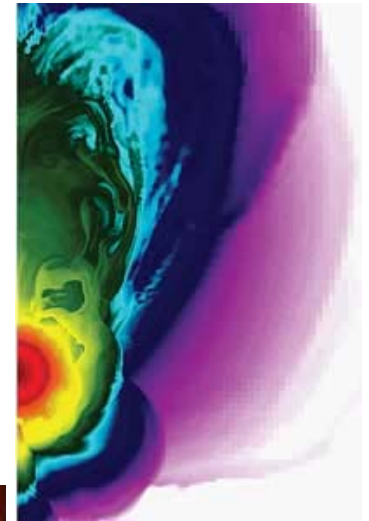
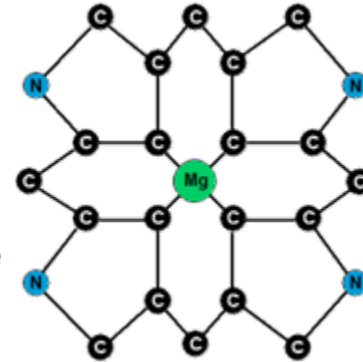




Large-Scale Capability Computing Is Addressing New Frontiers

INCITE Program at NERSC in 2004:

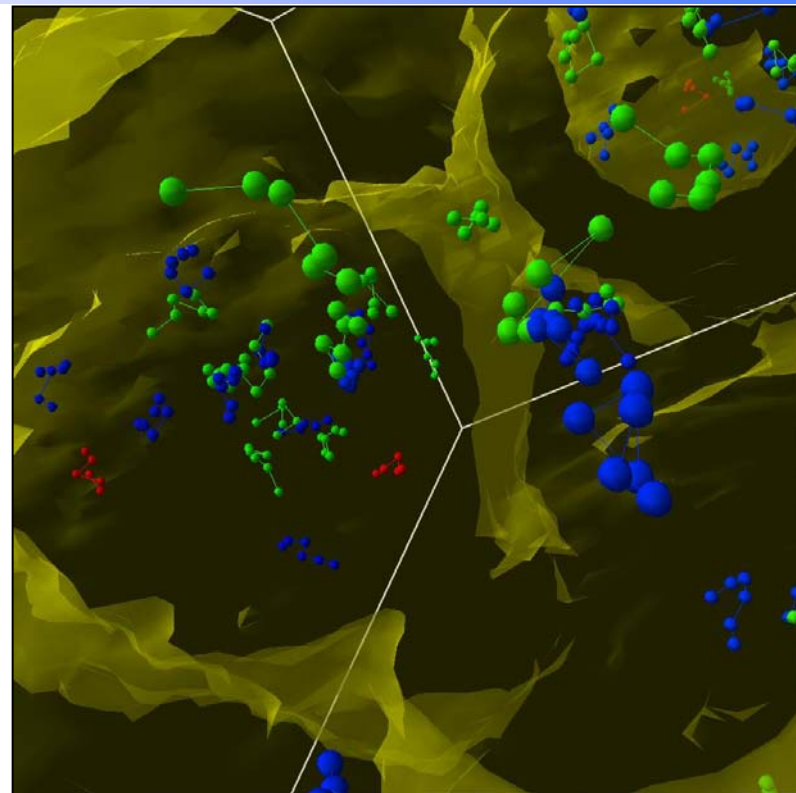
- Quantum Monte Carlo Study of Photosynthetic Centers; William Lester, Berkeley Lab
 - Largest QMC calculation to date anywhere (>600 electrons)
- Stellar explosions in three dimensions; Tomasz Plewa, University of Chicago
 - 3D simulations exhibit asymmetric explosions matching observed data
- Fluid Turbulence; P. K. Yeung, Georgia Institute of Technology
 - Largest DNS simulation in the U.S. on 2048**3 grid





INCITE: Quantum Monte Carlo Study of Photosynthesis

- **PI: William Lester and Graham Fleming, LBNL/UC Berkeley**
- **Goal: determine the ground to triplet-state energy difference of carotenoids present in photosynthesis**
- **Computation: Zori code for diffusion Quantum Monte Carlo, scaled to 4096 processors**
- **Results: most accurate values of the excitation and total energies of these biologically important systems; largest QMC calculation ever**

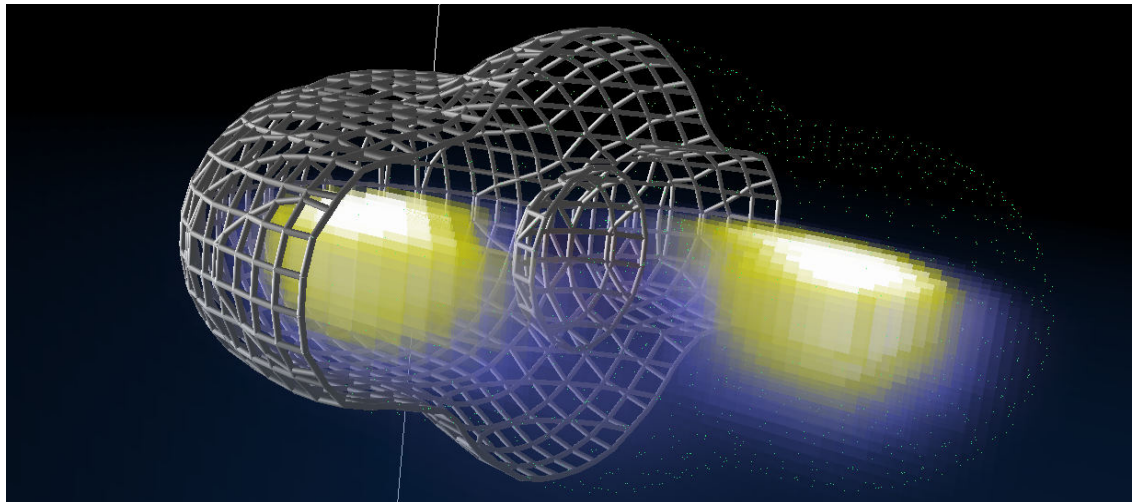


Imaginary time paths traversed by electrons in a photosynthetic system. The electrons are colored to make them distinct. The yellow isosurface shows the boundary of the molecular framework.



Black Hole Merger Simulations

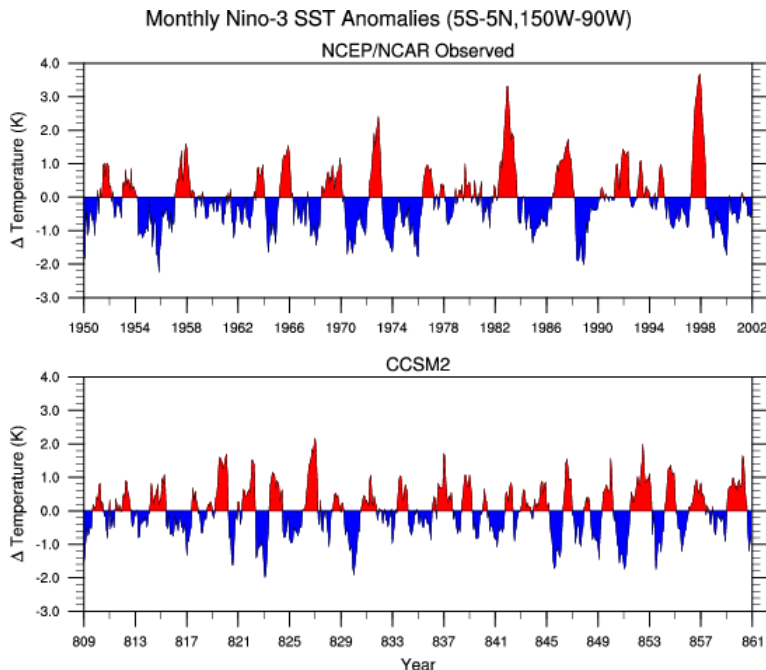
- *Ed Seidel, Gabrielle Allen, Denis Pollney, and Peter Diener, Max Planck Institute for Astrophysics; John Shalf, Lawrence Berkeley National Laboratory.*
 - Simulations of the spiraling coalescence of two black holes, a problem of particular importance for interpreting the gravitational wave signatures that will soon be seen by new laser interferometric detectors around the world.
 - First computed simulation of complete spiral of two binary black holes
-
- NERSC: 1.5 Tbytes of memory (nowhere else available at the time), visualization





1500 year climate simulation

- *Warren Washington and Jerry Meehl, National Center for Atmospheric Research; Bert Semtner, Naval Postgraduate School; John Weatherly, U.S. Army Cold Regions Research and Engineering Lab Laboratory.*

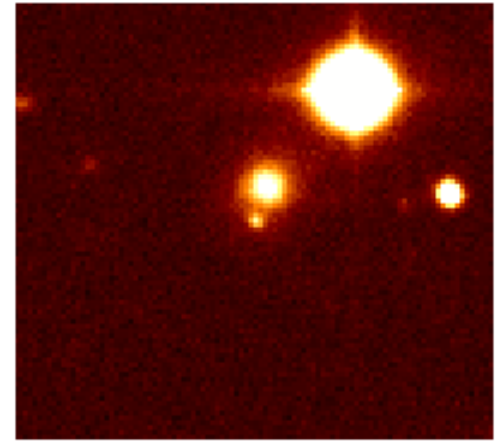


- 1,500-year simulation demonstrates the ability of the new Community Climate System Model (CCSM2) to produce a long-term, stable representation of the earth's climate.
- NERSC:
 - service and stability
 - special queue support
 - daily runs without impacting the rest of the workload



Nearby Supernova Factory

- Goal: Find and examine in detail up to 300 nearby Type Ia supernovae
 - More detailed sample against which older, distant supernovae can be compared
- Discovered 34 supernovae during first year of operation and now discovering 8-9 per month
- Previously a total of 130 supernovae were known
- First year: processed 250,000 images, archived
 - 6 TB of compressed data
- NERSC contribution:
 - high-speed data link
 - custom data pipeline software
 - NERSC's ability to store and process 50 gigabytes of data every night





Applications Scaling to Large Processor Counts

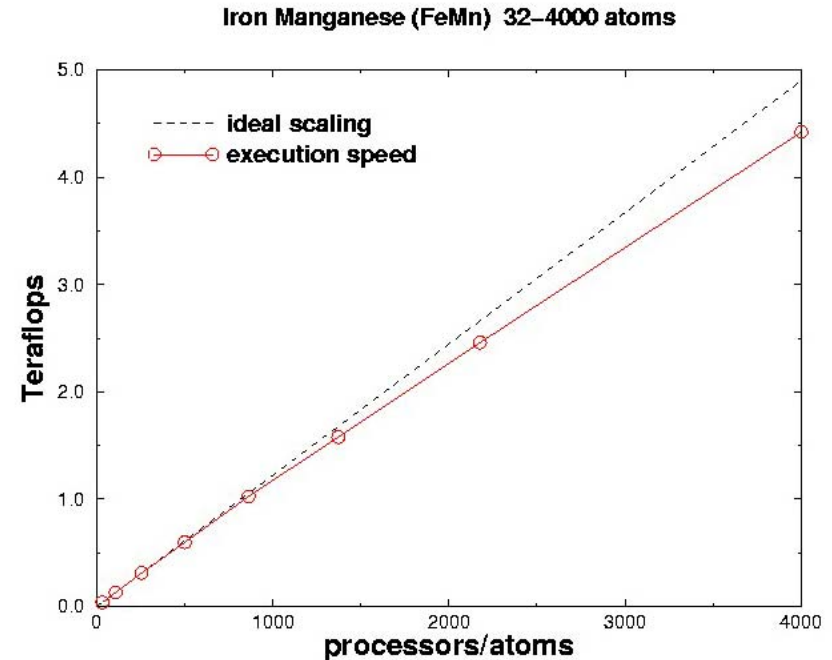
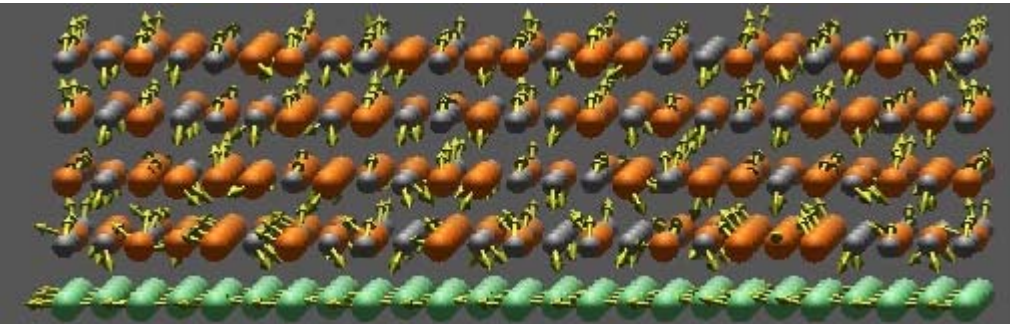
Multi-Teraflops Spin Dynamics Studies of the Magnetic Structure of FeMn and FeMn/Co Interfaces

Exchange bias, which involves the use of an antiferromagnetic (AFM) layer such as FeMn to pin the orientation of the magnetic moment of a proximate ferromagnetic (FM) layer such as Co, is of fundamental importance in magnetic multilayer storage and read head devices.

A larger simulation of 4000 atoms of FeMn ran at **4.42 Teraflops 4000 processors.**

(ORNL, Univ. of Tennessee, LBNL(NERSC) and PSC)

IPDPS03 A. Canning, B. Ujfalussy, T.C. Shulthess, X.-G. Zhang, W.A. Shelton, D.M.C. Nicholson, G.M. Stocks, Y. Wang, T. Dirks



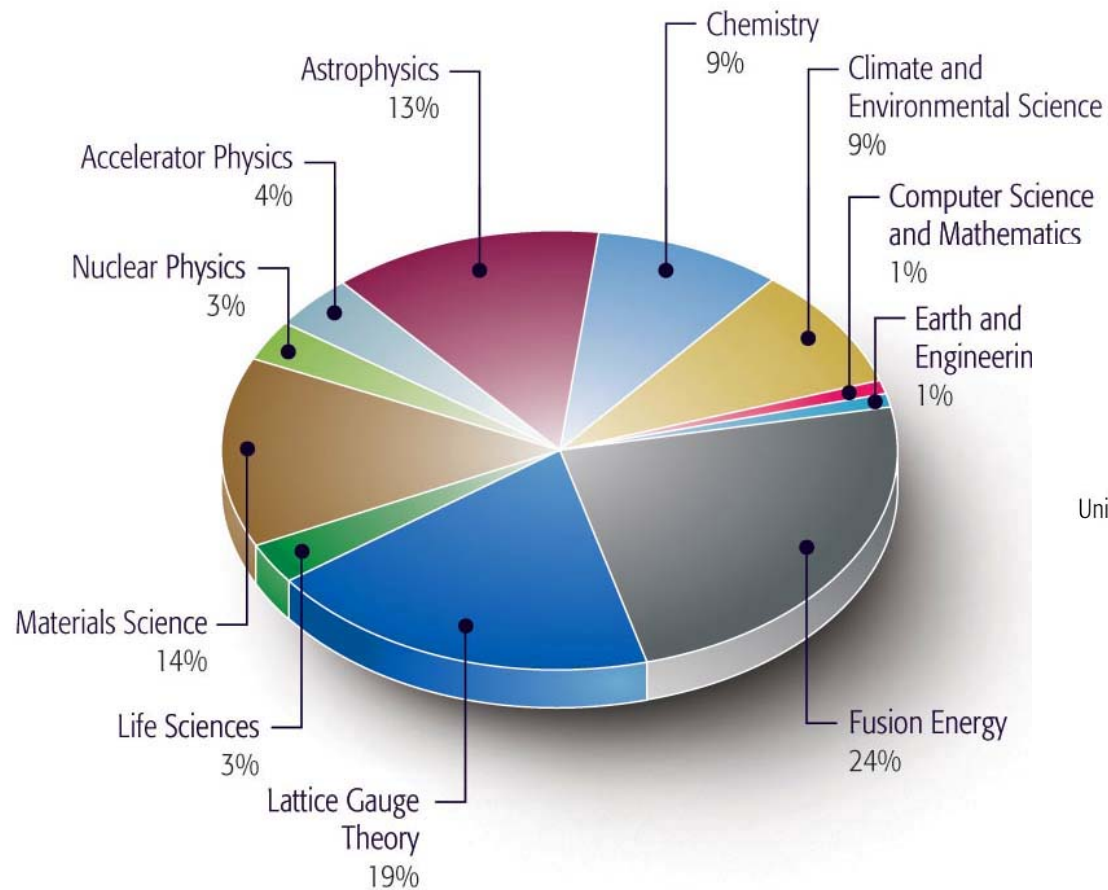
Section of an FeMn/Co (Iron Manganese/ Cobalt) interface showing the final configuration of the magnetic moments for five layers at the interface.

Shows a new magnetic structure which is different from the 3Q magnetic structure of pure FeMn.

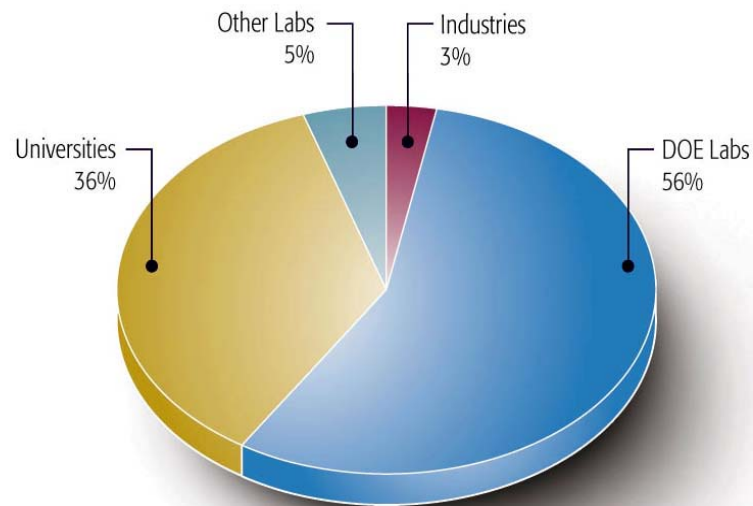


NERSC Supports A Diverse Science Community

NERSC Usage by Scientific Discipline,

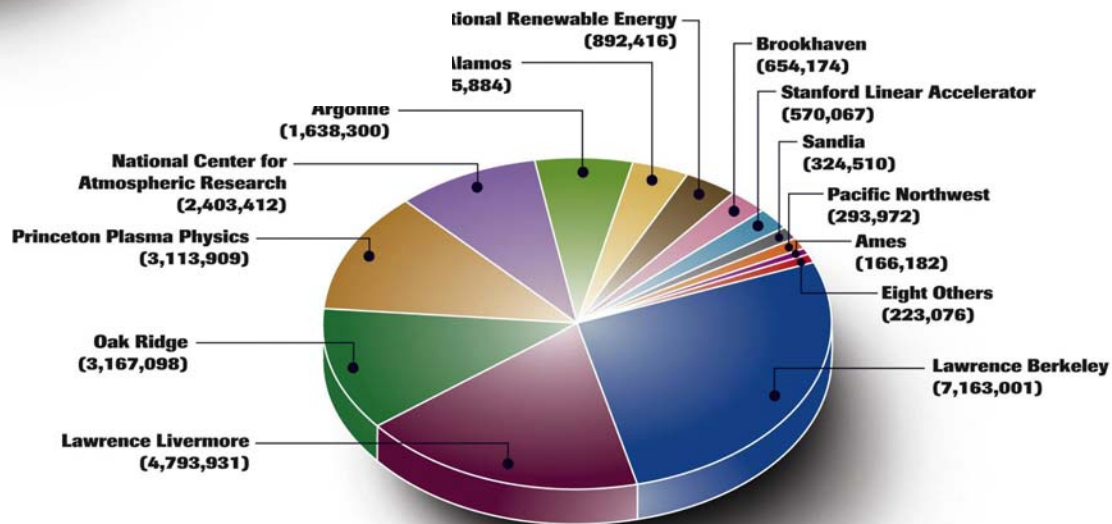
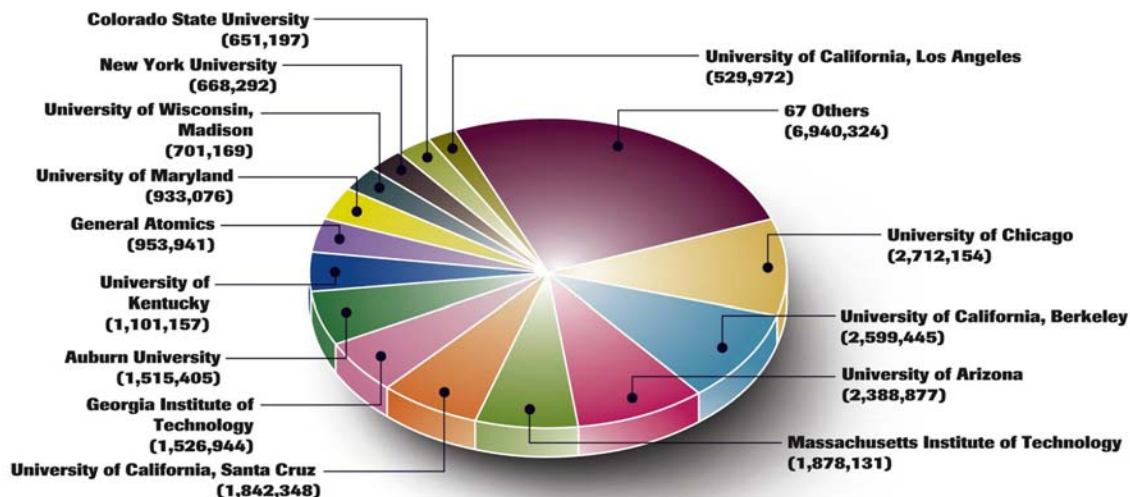


NERSC Usage by Institution Type,





Usage by Site, 2004 (processor hrs)



NERSC System Architecture


June 2005

Visualization Server – “escher”
SGI Onyx 3400 – 12 Processors
2 Infinite Reality 4 graphics pipes
24 Gigabyte Memory
5Terabytes Disk

Visualization Server – “Davinci”
SGI Altix – 8 Processors
48 Gigabyte Memory
3Terabytes Disk
(.5.62)

ETHERNET
10/100 Megabit

SYMBOLIC
~~MANIPULATION~~
SERVER



HPSS

HPSS

14 IBM SP servers
35 TB of cache disk

**8 STK robots, 44,000 tape slots,
24 - 200 GB drives, 60 - 20 GB drives
Max capacity 9 PB**

~~OC 48 – 2400 Mbps~~

ESnet

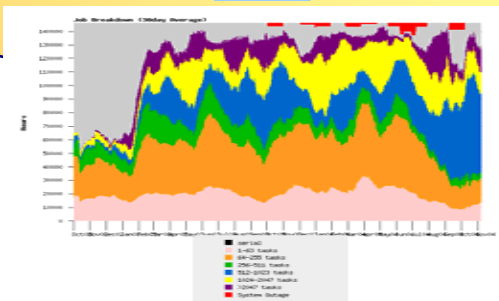
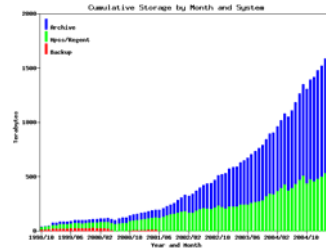
10 Gigabit, Gigabit Ethernet Jumbo

Testbeds and servers

NCS Cluster – “jacquard”
650 CPU Opteron/Infiniband 4X/12X
3.1 TF/ 1.2 TB memory
SSP - .41 Tflop/s
30 TB Disk
Ratio = (.4,10)

IBM SP
NERSC-3 – “Seaborg”
6,656 Processors (Peak 10 Tflop/s)
SSP – 1.35 Tflop/s
7.8 Terabyte Memory
55 Terabytes of Shared Disk
Ratio = (0.8,4.8)

PDSF
~800 processors
(Peak ~1.25 TFlop/s)
~1 TB of Memory
200 TB of Shared Disk
Gigabit and Fast Ethernet
Ratio = (0.8, 96)


$$\text{Ratio} = (\text{RAM Bytes per Flop}, \text{Disk Bytes per Flop})$$



Collaborative Acquisition

**DOING THINGS TOGETHER MEANS
(IN THIS CASE AT LEAST)
DOING THEM BETTER**



Coordination Is Useful

- **High End Computing Revitalization Task Force (HECRTF) – June 13-16, 2003**
 - “... benchmarking can be expensive for vendors...”
 - “...full-blown, competitive procurements are time-consuming and, hence, quite costly...”
 - “...it is paramount that ‘real’ benchmarks be used to categorize system performance. This is not the simplest of tasks even for short-term contracts”
 - “The current practice ... is to require vendors to provide performance results on some standard industry benchmarks and several scientific applications typical of those at the procuring site. Constructing these application benchmarks is a cost- and labor- intensive process, and responding to these solicitations is very costly for prospective vendors.”
 - “Recent successes with performance modeling suggest that it may be possible to accurately predict the performance of a future system, much larger than systems currently in use, on a scientific application much larger than any currently being run.”
 - “However, significant research is needed to make these methods usable ... Research is also needed to bolster capabilities to monitor and analyze the exploding volume of performance data that will be produced in future systems.”
- **Federal Plan for High-End Computing – May 10, 2004**
 - “The intent ... is to build teams that span agencies and increase visibility on issues critical to HEC procurement. ... expects that these projects will improve the information flow to assist in the prioritization of future HEC research, development, and engineering investments.”
 - “Moreover, coordinated procurement of HEC resources will provide more leverage in working with industry vendors to address the needs of the HEC applications communities.”
 - “... alternative approaches and planning strategies to carry out these activities. The current [method] allows for some evolutionary advances in high-end computing. However ... the current program will neither maintain U.S. leadership in the face of serious competition nor keep pace with the accelerating growth of demand for high-end computing resources to meet Federal agency needs.”
- **Getting Up to Speed – The Future of Supercomputing NAS Report – May 10, 2005**
 - “Performance modeling holds out of the hope of making a performance prediction of a system before it is procured, but currently modeling has only been done for a few codes by experts who have devoted a great deal of effort to understanding the code. To have a wider impact on the procurement process it will be necessary to simplify and automate the modeling process to make it accessible to nonexperts to use on more codes. Ultimately, performance modeling should become an integrative part of verification and validation for high-performance applications.”



What NERSC and HPCMP Have in Common

- Both NERSC and HPCMP centers are focused on facilitating scientific and engineering discover
 - Vast majority of resources devoted to “production computing”
 - Metric based programs
- Support a diverse mix of disciplines with users throughout the United States
- Acquire and Install leading edge computing systems with early production systems being the targets
- Need to integrate geographically remote sites
- Similar philosophy for evaluating systems
 - Measured performance with real application impact
- Facing the same technology challenges
 - The widening **gap** between application **performance** and peak performance of high-end computing systems
 - The recent emergence of **large, multidisciplinary** computational science **teams** in the DOE research community
 - The **flood of** scientific **data** from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows



Some Difference between NERSC and HPCMP

- NERSC has no restricted, sensitive or classified work
- NERSC does a major new system every three years with smaller, more focused systems in between.
- The NERSC budget is less than 15% of HPCMP
- 90% of NERSC usage is by codes that are developed and maintained by the projects scientists
- Probably more if we go deeper



TI-06/NERSC-5 Collaboration



NERSC-5 Goals

- **Support the entire NERSC Workload**
- **Significant increase over the combined NERSC-3, NCS, NCSb using measured performance criteria**
 - Expected to significantly increase computational time for NERSC users in the 2007 Allocation Year
 - Dec 1, 2006 – November 30, 2007
 - Have full impact for AY 2008
 - Can arrive in FY 2006
- **Integrate with the NERSC environment**
- **Sustained System Performance over 3 years**
- **System Balance**
 - Aggregate memory
 - Memory per CPU?
 - Global usable disk storage
 - IO Bandwidth and Latency
 - Global
 - Per CPU
 - TB storage
 - Network BW



NERSC-5

- NERSC uses a “Best Value” approach
- Requirements derived from the NERSC User Group “Greenbook” a comprehensive set of scientific requirements created every three years.
- We expect systems that have all the features of good, integrated early production systems
 - PERVU Focus
 - A “holistic” evaluation methodology for large systems
 - Performance
 - Many ways to determine this – some better than others
 - » Application benchmarks
 - » Linpack, NPBs, etc
 - » Sustained System Performance (SSP) tests
 - Effectiveness
 - Effective System Performance (ESP) Test
 - Reliability
 - Looking for new ways to proactively assess
 - Variation
 - CoV and other methods
 - Usability
 - A relative metric of usability
- This discussion will focus on benchmarking and evaluation not all the features



Overview of Collaborative Procurement

- In 2004, NERSC/HPCMP identified the potential for working together
- TI-06 and NERSC RFP within a couple of months of each other.
 - Both use a mix of application codes and kernels
 - Some similarities in style of procurement
 - NERSC 5 system is scheduled for delivery 6-9 months after TI-06 systems
- Expertise in both organizations
 - Computer Architecture Evaluation
 - Application and kernel benchmarking
 - Composite benchmarks
 - Modeling
 - System Management and Operation
 - User Support
- NERSC staff attended the TI-06 Performance Group Meetings
- TI-06 staff (Roy Campbell, Bill Ward, Dave Koester) visited NERSC

Decided it was worthwhile to work together to see what could be done



Application Benchmarks

- **NERSC and HPCMP developed up with their own list of benchmark candidates**
 - **NERSC derives its application benchmark from the actual workload**
 - **Starts with 15-20 candidates from user community**
 - **Selection of benchmarks use several considerations**
 - **Representative of the workload**
 - **Represent different algorithms and methods**
 - **Are portable to candidate architectures with limited effort**
 - **Work in a repeatable and testable manner**
 - **Are tractable for a non-expert to understand**
 - **Can be instrumented**
 - **Can be distributed at least to potential bidders**
 - **Arrived at suite of 8 for NERSC-5**
- **Once candidates were narrowed down NERSC and HPCMP staff reviewed the candidates**



NERSC-5 Application Summary

| Application | Science Area | Basic Algorithm | Language | Library Use | Originating Organization |
|-------------|-------------------------|---------------------------|------------|-------------|---|
| CAM3 | Climate (BER) | Navier-Stokes CFD | FORTRAN 90 | netCDF | NCAR |
| GAMESS | Chemistry (BES) | DFT | FORTRAN 90 | DDI, BLAS | Iowa State, Ames Laboratory |
| GTC | Fusion (FES) | Particle-in-cell | FORTRAN 90 | FFT(opt) | PPPL |
| MADbench | Astrophysics (HEP & NP) | Power Spectrum Estimation | C | Scalapack | LBNL |
| MILC | QCD (NP) | Sparse Conjugate gradient | C | none | Wide collaboration |
| PARATEC | Materials (BES) | 3D FFT | FORTRAN 90 | Scalapack | LBNL and UCB |
| PMEMD | Life Science (BER) | Particle Mesh Ewald | FORTRAN 90 | none | University of North Carolina, Chapel Hill |



TI-06 Application Benchmark Codes

Aero – Aeroelasticity CFD code

(Fortran, serial vector, 15,000 lines of code)

AVUS - CFD calculations on unstructured grids

(MPI)

CTH – Shock physics code

(~43% Fortran/~57% C, MPI, 436,000 Lines of code)

GAMESS – Quantum chemistry code

(Fortran, MPI, 330,000 Lines of code)

HYCOM – Ocean circulation modeling code

(Fortran, MPI, 31,000 Lines of code)

LAMMPS – Molecular Dynamics code for micro and macro scale

(C++, Fortran)

OOCore – Out-of-core solver

(Fortran, MPI, 39,000 Lines of code)

Overflow-2 – CFD code originally developed by NASA

(MPI with OpenMP)

WRF – Multi-Agency mesoscale atmospheric modeling code

(Fortran and C, MPI, 100,000 Lines of code)



Looked for Commonality

- Realized the NERSC and HPCMP workloads had significant differences and only a few common areas
 - HPCMP
 - More structural codes – very little at NERSC
 - More external CFD
 - NERSC
 - More astrophysics, QCD, fusion and life sciences
 - Common areas
 - Chemistry
 - GAMESS has a major code at both sites
 - Weather/climate
 - DOE does not do weather research, so CCSM is more appropriate of our workload
 - There was not enough time to consider CSM for the T1 06 suite
 - Maybe a candidate area for next time
 - Computational Fluid Dynamics
 - NERSC has focus on AMR CFD that focus's on combustion studies
- The only practical common benchmark was GAMESS for this period
 - NERSC adopted the HPCMP run rules so vendors only have to do this once
 - Using the same scaling and data sets



Kernel Benchmarks

- **NERSC has a lot of experience with the NAS Parallel Benchmarks and our LBNL performance research collaborators have developed some new tests**
 - **Processor: NAS Parallel Benchmarks (NPB)**
 - Serial: NPB 2.3 Class B
 - Parallel: NPB 2.4 Class D at 64 256 processors
 - **Memory**
 - Streams
 - APEX Mp – serial
 - **Interconnect**
 - PingPong
 - APEX Mpparallel
 - **I/O benchmark**
 - Pioraw based benchmark
 - **Full configuration test**
 - global FFT or reduction operation
- **HPCMP has analogous but different set**
 - CPUbench
 - MEMbench
 - NETbench
 - OSBench
 - SPIOBench

Both sides want to explore modeling so that is where overlap made sense



Modeling

- DOE funds both the two major modeling efforts
 - PMAC (Alan Snavley and Laura Carrington) at SDSC
 - Adolfo Hoisie at LANL
- HPCMP is funding Alan to model all the TI-06 codes
- NERSC, with the support of DOE/MICS, engage both groups to model some of the NERSC Applications and compare experiences
 - The effort and experiences of a non-originator to use the modeling methods
 - Comparison of the two most often methods approaches
 - Accuracy and feasibility of modeling for procurement
 - Comparison of predicted vs actual systems
- Both HPCMP and NERSC are using the exact same kernels needed to give models information
 - These need to be run to establish parameters for performance modeling of application codes.
 - Memory Test: Membench
 - Communication tests: Netbench
 - broadcast
 - allreduce
 - PingPong
 - PingPing
 - I/O test: StreamIO
 - NERSC is replacing its standard IO test with this code



Observations

- **Sharing experiences and expertise is very beneficial**
- **Vendors generally pleased to hear there is overlap**
 - Not clear they believe it until they see it however
- **Application benchmarks derived from the science workload at each site have limited overlap**
 - **Therefore limited opportunity to consolidate**
 - Still useful to try since it is little effort to evaluate and saves 8-15% effort for each success
- **Collaborative micro/synthetic benchmarks may leave more opportunity**
- **Modeling only helps to a degrees**
 - **Saves vendors from having to run applications but it is at least as much effort for procurement teams**
 - Have to trace and model the codes
 - Have to have full application codes ready for validation and acceptance
 - Introduce more risk for sites
- **Collaborative composite benchmark methods are promising areas**
- **Different sites have different procurement schedules, but in general more sites should coordinate efforts**



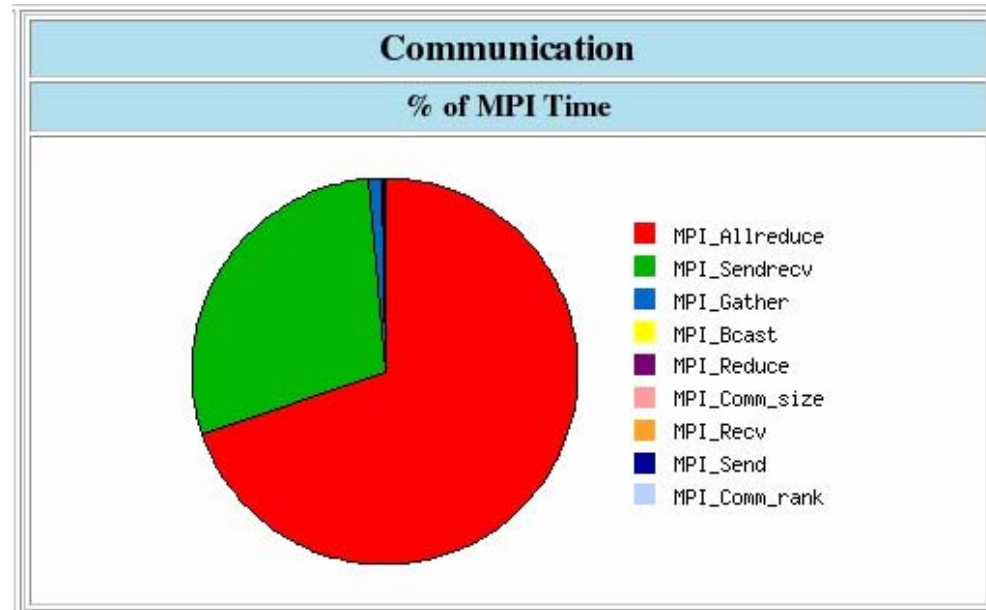
Composite Benchmarks and Metrics

Possible future collaborative
areas



Performance Analysis

- NERSC provides performance data with all its benchmarks
- IPM – a system independent performance collection tool to instrument codes
- HPCMP is exploring the use of IPM on several of your systems

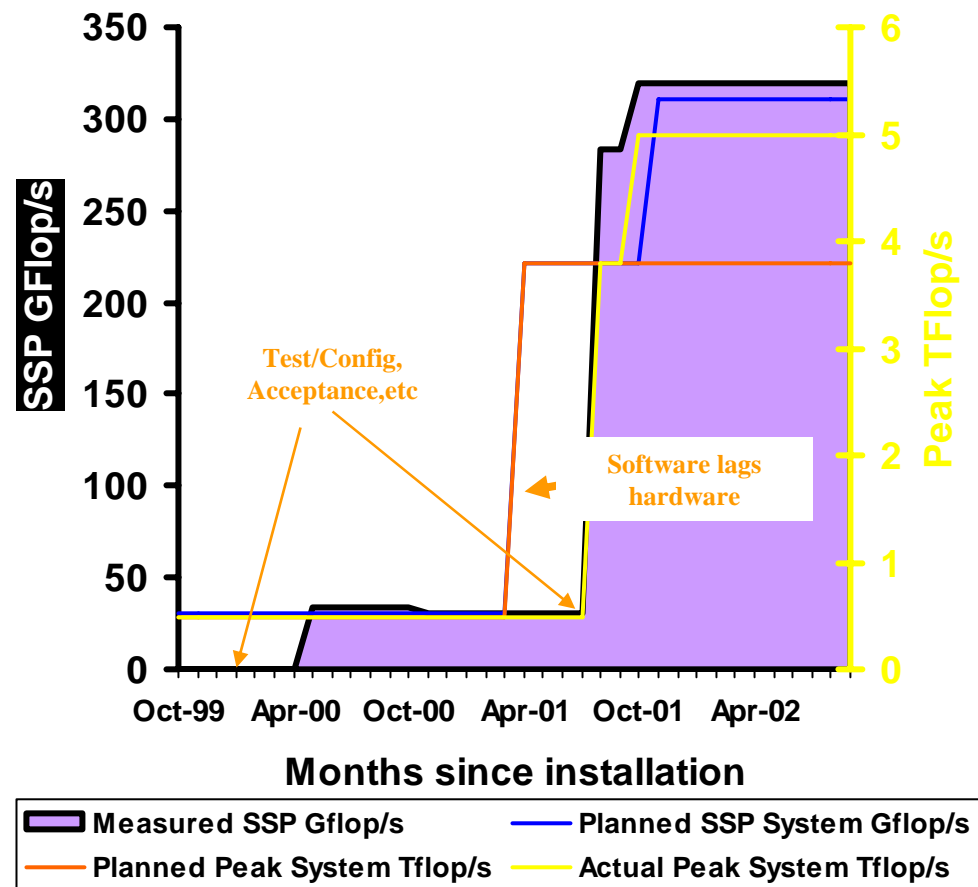




Sustained System Performance (SSP) Test

- NERSC focuses on the area under the measured curve
 - SSP is responsible for assuring delivered performance
 - SSP is conservative so most applications do better
 - To achieve the required performance, NERSC-3 has a 22% higher peak performance than planned
 - The higher final result benefits the community for the long term

Peak vs SSP



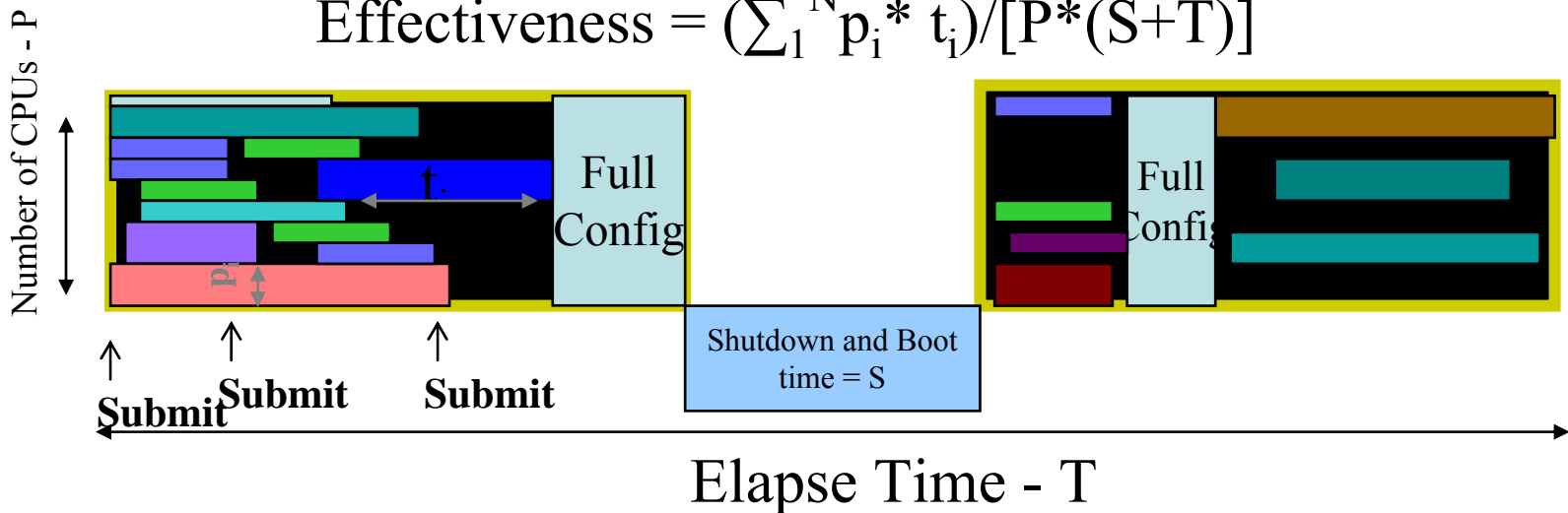
$$\text{SSP} = \text{Measured Performance} * \text{Time}$$



Effective System Performance (ESP) Test

- Test uses a mix of NERSC test codes, that run in a random order, testing standard system scheduling.
 - There are also Full Configuration codes, I/O tests and typical System Administration activities.
- Independent of hardware and compiler optimization improvements
- The test measures both how much and how often the system can do scientific work

$$\text{Effectiveness} = (\sum_1^N p_i * t_i) / [P * (S + T)]$$





Variation

- Large Scale multiprocessors can exhibit large variation in performance as measured by run time
 - Even on well managed systems where jobs run on dedicated nodes
 - It takes tremendous effort to get the state of “well managed”
- Performance variation causes problems for users
 - Loss of user productivity
 - Jobs abort when limits exceeded
 - Run time estimates have to be conservative in order to accommodate variation
 - With accurate estimates more progress can be made
 - Loss of system productivity
 - Less work can go through the system
 - Less reliable estimates of usage make job scheduling less effective
- Questions
 - How to know if a system has high variation
 - What can be done to minimize variation



IO Benchmarking

- Future storage system will be more complex
- Diverse I/O workloads need flexible tests
 - Large I/O
 - POSIX I/O
 - Parallel I/O
 - Many file I/O
 - Metadata
 - Manipulation of directory information
 - Compiling large sets of files



Berkeley Institute for Performance Studies

- Application evaluation on vector processors
- Architectural probes for alternative architectures
- APEX: Application Performance Characterization Benchmarking
- BeBop: Berkeley Benchmarking and Optimization Group
- LAPACK: Linear Algebra Package
- Modern Vector Architecture
- PERC: Performance Engineering Research Center
- Top500
- ViVA: Virtual Vector Architectures



NERSC -2006-2010

Opportunities for closer relationships



Science-Driven Computing Strategy 2006 -2010





Science-Driven Systems

- **Science-Driven Systems**
 - Balanced and timely introduction of best new technology for complete computational systems (computing, storage, networking, analytics)
 - Engage and work directly with vendors in addressing the SC requirements in their roadmaps
 - Collaborate with DOE labs and other sites in technology evaluation and introduction
- **Science-Driven Services**
 - Provide the entire range of services from high-quality operations to direct scientific support
 - Enable a broad range of scientists to effectively use NERSC in their research
 - Concentrate on resources for scaling to large numbers of processors, and for supporting multidisciplinary computational science teams
- **Science-Driven Analytics**
 - Provide architectural and systems enhancements and services to more closely integrate computational and storage resources
 - Provide scientists with new tools to effectively manipulate, visualize and analyze the huge data sets from both simulations and experiments



What is Analytics?

- **Science of reasoning**
 - Generate insight and understanding from large, complex, disparate, sometimes conflicting data
- **Visual analytics:**
 - Science of reasoning facilitated by visual interfaces
- **Why visual?**
 - High bandwidth through human visual system
 - Better leverage human reasoning, knowledge, intuition and judgment
- **Intersection of:**
 - Visualization, analysis, scientific data management, human-computer interfaces, cognitive science, statistical analysis, reasoning, ...
- **Solutions are domain-specific combinations of above technologies**



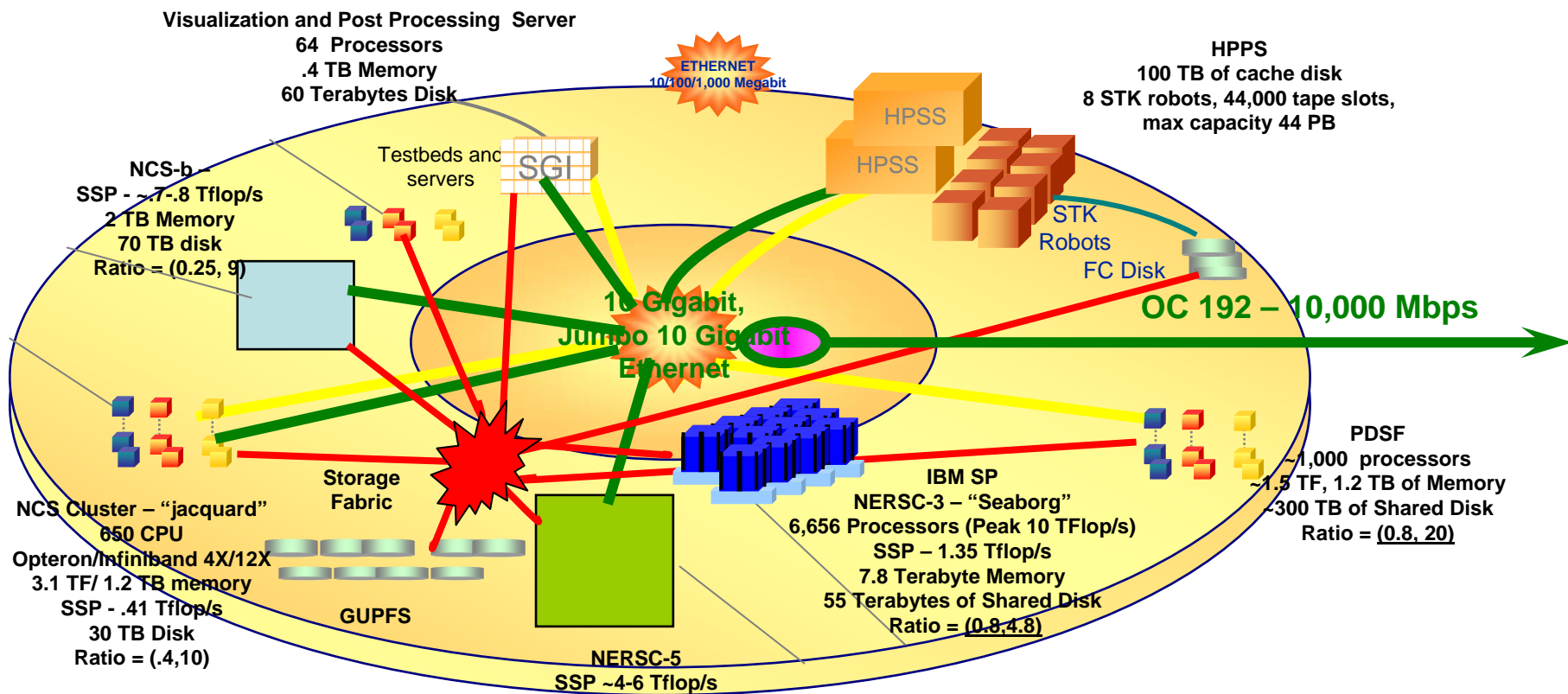
Why Analytics?

- All sciences need to find, access, and store and understand information
 - Data is the limiting or the enabling factor for a wide range of sciences
- Synthesize information and derive insight from massive, dynamic, ambiguous and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible and understandable findings
- Effectively communicate findings
- In some sciences, the data management (and analysis) challenge already exceeds the compute-power challenge in required resources
- The ability to deal with a tidal wave of information will distinguish the most successful scientific, commercial, and national security endeavors



NERSC's Analytics Strategy

- **Broad strategic program objectives:**
 - Clear picture of user needs
 - Leverage existing and provide new visualization and analysis capabilities
 - Enhance data management infrastructure
 - Enhance distributed computing infrastructure
 - Realizing analytics: support for the NERSC user community



Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)



Science Driven System Architecture



Federal Plan for High-End Computing – May 10, 2004

- “The high-end marketplace today is not producing machines with the required capabilities to satisfy the most demanding scientific applications. Where there is substantial overlap between commercial computing needs and scientific needs, vendors are supplying products with astounding performance. However, where scientific or defense needs do not overlap substantially with commercial IT, the product space is lacking.”



Applications and Algorithms Matrix

| Science areas | Multi-physics, Multi-scale | Dense linear algebra | Sparse linear algebra | FFT's | AMR | Data Intensive |
|---------------|---------------------------------|----------------------------------|--------------------------------|----------------------------------|---------------------------------|---------------------------------|
| Nanoscience | General purpose balanced system | High speed CPU, high Flop/s rate | High performance memory system | Bisection interconnect bandwidth | Irregular data and control flow | Storage, Network Infrastructure |
| Climate | | | | | | |
| Chemistry | | | | | | |
| Fusion | | | | | | |
| Combustion | | | | | | |
| Astrophysics | | | | | | |
| Biology | | | | | | |
| Nuclear | | | | | | |



Science Driven System Architectures Goals

- Broadest, large-scale application base runs very well on SDSA solutions with excellent *sustained* performance per dollar
- Even applications that do well on specialized architectures could perform near optimal on a SDSA Architectures





Science Driven System Architecture Goals

- Collaboration between scientists and computer vendors on science driven system architecture is the path to continued improvement in application performance
- Create systems that best serve the entire science community
- Vendors do not design to the future benchmarks, they design to the past benchmarks. Hence, just relying on procurement benchmarks is not getting us to where we need to be
 - Vendors are not knowledgeable in current and future algorithmic methods.
 - When SDSA started, system designers were working with algorithms that were 10 years old
 - Did not consider sparse matrix methods of 3D FFTs in design of CPUs
- Active collaboration with scientific application community and the computer science community will address many of these issues
- Early objectives:
 - ViVA-2 architecture development – Power6 scientific application accelerator
 - Additional investigation with other architectures
 - Lower interconnect latency and large spanning
- Long-term objectives:
 - Integrate lessons of the large scale systems, such as the Blue Gene/L and HPCS experiments, with other technologies, into a hybrid system for petascale computing.
- SDSA applies to all aspect – not just parallel computing
 - Facility Wide File Systems



SCDA Results : LBNL Blue Planet

2002: Berkeley Lab launches science-driven architecture process

2003: Multiple design discussions, reviews with scientists and computer architects at Berkeley Lab, LLNL, IBM

“IBM has already adopted the concepts of ‘Science Driven Architecture Design’ in redesigning the Power 5/6 node. We will continue the Science Driven Design approach...”

Nicholas Donofrio, Senior Vice President, IBM

2004: IBM incorporates Blue Planet node design and enhanced interconnect in product roadmap and will deliver first implementation to ASCI program

“The Blue Planet node conceived by NERSC and IBM [...] features high internal bandwidth essential for successful scientific computing. LLNL elected early in 2004 to modify its contract with IBM to use this node as the building block of its 100 TF Purple system. This represents a singular benefit to LLNL and the ASC program, and LLNL is indebted to LBNL for this effort.”

Dona Crawford, Associate Director for Computation, LLNL

2005: Release of performance data for NERSC 5 applications, work on ViVA-1 and ViVA-2, beginning exploration of interconnect alternatives



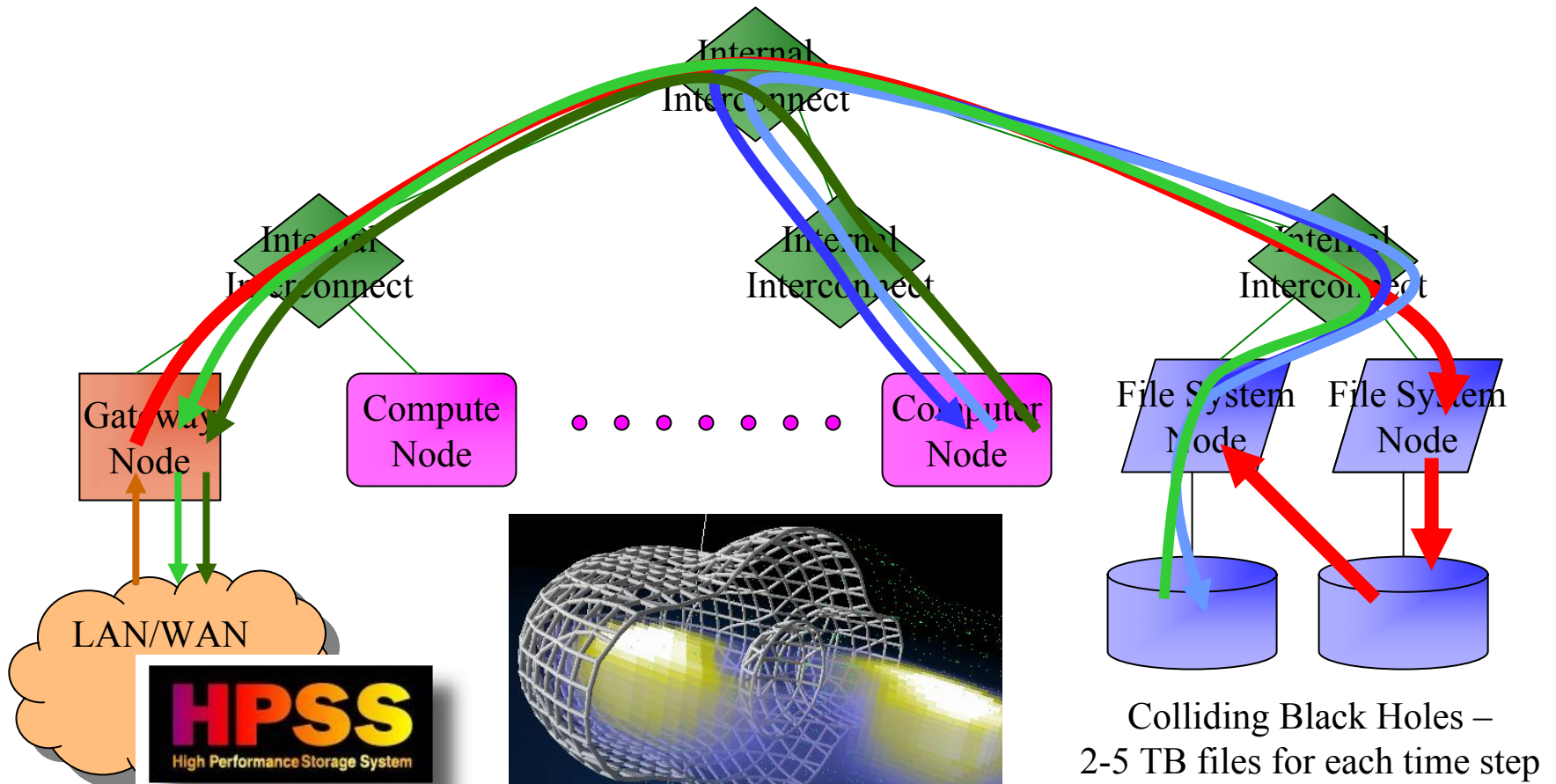
Facility Wide File Systems

Global File Systems



Data Divergence Problem

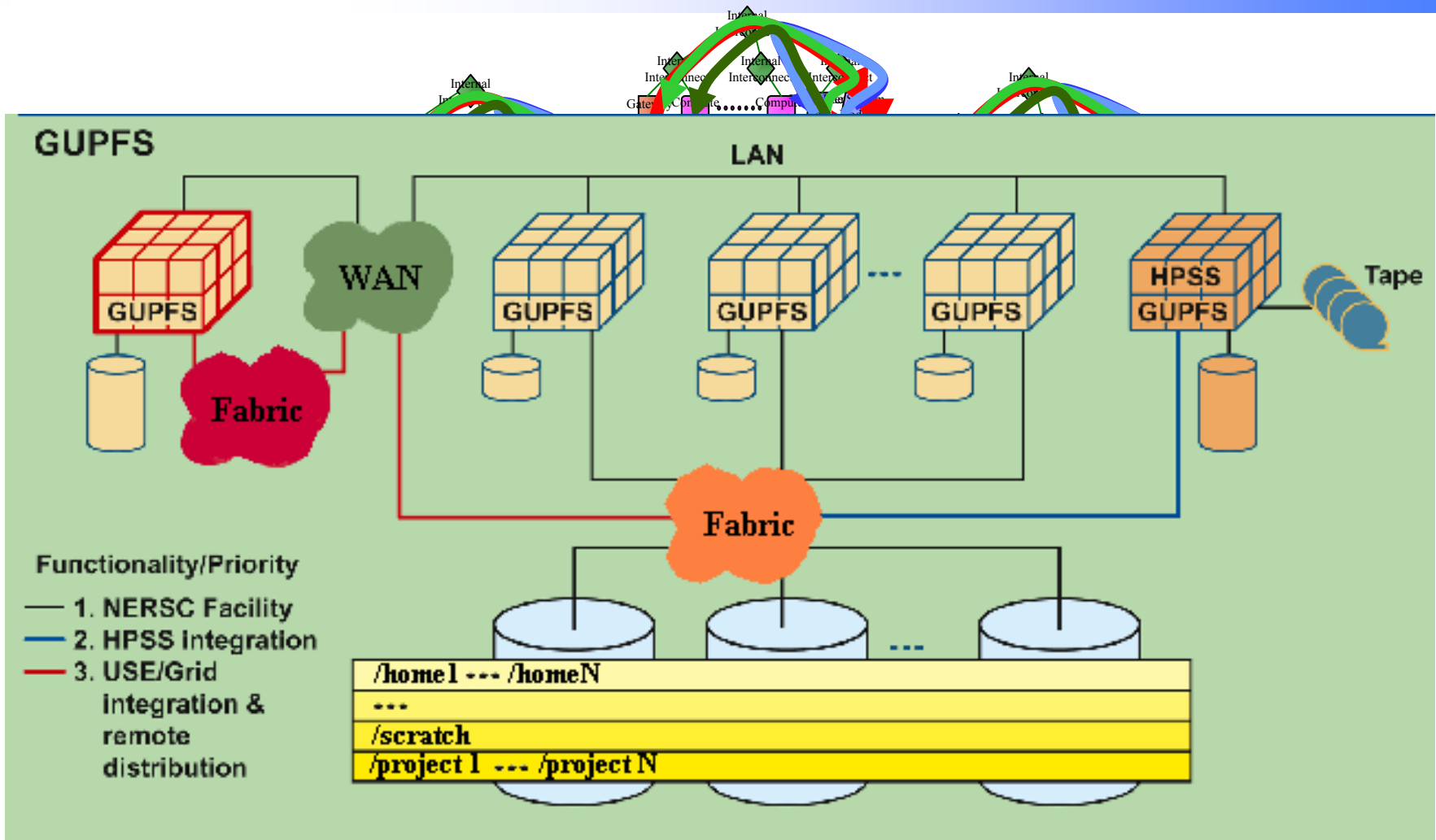
The memory divergence problem is masking the data divergence problem



Colliding Black Holes –
2-5 TB files for each time step



Facility Wide File System





Facility Wide File System Deployment

- **FY 05: initial production deployment with shared file system functionality and features**
 - Minimal 20 TB usable end user storage and 1 GB/s bandwidth for streaming I/O
 - Storage and servers external to all client systems
 - Distributed over a 10 Gigabit Ethernet infrastructure
 - Single file system instance providing file and data sharing among multiple client systems
 - Both large and small files expected
 - Not a scratch or a home file system
 - Focus on function first, then performance
- **Initial clients are intended to be:**
 - Seaborg, IBM SP running AIX 5.2
 - Jacquard, LNXI Opteron System running SLES 9
 - Da Vinci, SGI Altix running SLES 9
 - NCSb
 - (Possibly) PDSF IA32 Linux cluster running RHEL



Summary

- **HPCMP is one of the leaders in how to evaluate and select HPC systems. The program is also a key leader in implementing the Revitalization of HEC**
- **The effort on TI-06/NERSC-5 collaboration has been effective**
 - **In addition to the commonality of benchmarks, the investigation of modelling's ability to take a larger role in procurements should help the entire community.**
- **There are further opportunities for NERSC and HPCMP to work together and make improvements in HPC**
 - **Expand benchmarking and analysis**
 - **Modeling**
 - **Composite Benchmarks**
 - **Science Driven System Architecture**
 - **Science Driven Analytics**



Discussion

